Learning with Spectral Kernels and Heavy-Tailed Data[*]

Michael W. Mahoney [†] Hariharan Narayanan [‡]

Abstract

Two ubiquitous aspects of large-scale data analysis are that the data often have heavy-tailed properties and that diffusion-based or spectral-based methods are often used to identify and extract structure of interest. Perhaps surprisingly, popular distribution-independent methods such as those based on the VC dimension fail to provide nontrivial results for even simple learning problems such as binary classification in these two settings. In this paper, we develop distribution-dependent learning methods that can be used to provide dimension-independent sample complexity bounds for the binary classification problem in these two popular settings. In particular, we provide bounds on the sample complexity of maximum margin classifiers when the magnitude of the entries in the feature vector decays according to a power law and also when learning is performed with the so-called Diffusion Maps kernel. Both of these results rely on bounding the annealed entropy of gap-tolerant classifiers in a Hilbert space. We provide such a bound, and we demonstrate that our proof technique generalizes to the case when the margin is measured with respect to more general Banach space norms. The latter result is of potential interest in cases where modeling the relationship between data elements as a dot product in a Hilbert space is too restrictive.

1 Introduction

Two ubiquitous aspects of large-scale data analysis are that the data often have heavy-tailed properties and that diffusion-based or spectral-based methods are often used to identify and extract structure of interest. In the absence of strong assumptions on the data, popular distribution-independent methods such as those based on the VC dimension fail to provide nontrivial results for even simple learning problems such as binary classification in these two settings. At root, the reason is that in both of these situations the data are formally very high dimensional and that (without additional regularity assumptions on the data) there may be a small number of "very outlying" data points. In this paper, we develop distribution-dependent learning methods that can be used to provide dimension-independent sample complexity bounds for the maximum margin version of the binary classification problem in these two popular settings. In both cases, we are able to obtain nearly optimal linear classification hyperplanes since the distribution-dependent tools we employ are able to control the aggregate effect of the "outlying" data points. In particular, our results will hold even though the data may be infinite-dimensional and unbounded.

1.1 Overview of the problems

Spectral-based kernels have received a great deal of attention recently in machine learning for data classification, regression, and exploratory data analysis via dimensionality reduction [25]. Consider, for example, Laplacian Eigenmaps [2] and the related Diffusion Maps [6]. Given a

[*] This version of this paper is substantially revised and extended relative to the first version of 24 June 2009.

[†] Department of Mathematics, Stanford University, Stanford, CA 94305, mmahoney@cs.stanford.edu.

[‡] Laboratory for Information and Decision Systems, MIT, Cambridge, MA 02139, har@mit.edu.

graph $G = (V, E)$ (where this graph could be constructed from the data represented as feature vectors, as is common in machine learning, or it could simply be a natural representation of a large social or information network, as is more common in other areas of data analysis), let $f_0, f_1, \ldots, f_n$ be the eigenfunctions of the normalized Laplacian of G and let $l_0, l_1, \ldots, l_n$ be the corresponding eigenvalues. Then, the Diffusion Map is the following feature map

$$\Phi : v \mapsto (l_0^k f_0(v), \ldots, l_n^k f_n(v)),$$

and Laplacian Eigenmaps is the special case when $k = 0$. In this case, the support of the data distribution is unbounded as the size of the graph increases; the VC dimension of hyperplane classifiers is $\Theta(n)$; and thus existing results do not give dimension-independent sample complexity bounds for classification by Empirical Risk Minimization (ERM). Moreover, it is possible (and indeed quite common in certain applications) that on some vertices v the eigenfunctions fluctuate wildly—even on special classes of graphs, such as random graphs $G(n, p)$, a non-trivial uniform upper bound stronger than $O(n)$ on $\|\Phi(v)\|$ over all vertices v does not appear to be known.[1] Even for maximum margin or so-called "gap-tolerant" classifiers, defined precisely in Section 2 and which are easier to learn than ordinary linear hyperplane classifiers, the existing bounds of Vapnik are not independent of the number n of nodes.[2]

A similar problem arises in the seemingly very-different situation that the data exhibit heavy-tailed or power-law behavior. Heavy-tailed distributions are probability distributions with tails that are not exponentially bounded [24, 5]. Such distributions can arise via several mechanisms, and they are ubiquitous in applications [5]. For example, graphs in which the degree sequence decays according to a power law have received a great deal of attention recently. Relatedly, such diverse phenomenon as the distribution of packet transmission rates over the internet, the frequency of word use in common text, the populations of cities, the intensities of earthquakes, and the sizes of power outages all have heavy-tailed behavior. Although it is common to normalize or preprocess the data to remove the extreme variability in order to apply common data analysis and machine learning algorithms, such extreme variability is a fundamental property of the data in many of these application domains.

There are a number of ways to formalize the notion of heavy-tailed behavior for the classification problems we will consider, and in this paper we will consider the case where the *magnitude* of the entries decays according to a power law. (Note, though, that in Appendix A, we will, for completeness, consider the case in which the probability that an entry is nonzero decays in a heavy-tailed manner.) That is, if

$$\Phi : v \mapsto (\phi_0(v), \ldots, \phi_n(v))$$

represents the feature map, then $\phi_i(v) \leq C i^{-\alpha}$ for some absolute constant $C > 0$, with $\alpha > 1$. As in the case with spectral kernels, in this heavy-tailed situation, the support of the data distribution is unbounded as the size of the graph increases, and the VC dimension of hyperplane classifiers

[1] It should be noted that, while potentially problematic for what we are discussing in this paper, such eigenvector localization often has a natural interpretation in terms of the processes generating the data and can be useful in many data analysis applications. For example, it might correspond to a high degree node or an articulation point between two clusters in a large informatics graph [17, 16, 18]; or it might correspond to DNA single-nucleotide polymorphisms that are particularly discriminative in simple models that are chosen for computational rather than statistical reasons [19, 23].

[2] VC theory provides an upper bound of $O\left((n/\Delta)^2\right)$ on the VC dimension of gap-tolerant classifiers applied to the Diffusion Map feature space corresponding to a graph with n nodes. (Recall that by Lemma 2 below, the VC dimension of the space of gap-tolerant classifiers corresponding to a margin Δ, applied to a ball of radius R is $\sim (R/\Delta)^2$.) Of course, although this bound is quadratic in the number of nodes, VC theory for ordinary linear classifiers gives an $O(n)$ bound.

is $\Theta(n)$. Moreover, although there are a small number of "most important" features, they do not "capture" most of the "information" of the data. Thus, when calculating the sample complexity for a classification task for data in which the feature vector has heavy-tailed properties, bounds that do not take into account the distribution are likely to be very weak.

In this paper, we develop distribution-dependent bounds for problems in these two settings. Clearly, these results are of interest since VC-based arguments fail to provide nontrivial bounds in these two settings, in spite of ubiquity of data with heavy-tailed properties and the widespread use of spectral-based kernels in many applications. More generally, however, these results are of interest since the distribution-dependent bounds underlying them provide insight into how better to deal with heterogeneous data with more realistic noise properties.

1.2 Summary of our main results

Our first main result provides bounds on classifying data whose magnitude decays in a heavy-tailed manner. In particular, in the following theorem we show that if the weight of the i^{th} coordinate of random data point is less than $Ci^{-\alpha}$ for some $C > 0, \alpha > 1$, then the number of samples needed before a maximum-margin classifier is approximately optimal with high probability is independent of the number of features.

Theorem 1 (Heavy-Tailed Data) *Let the data be heavy-tailed in that the feature vector:*

$$\Phi : v \mapsto (\phi_1(v), \ldots, \phi_n(v)),$$

satisfy $|\phi_i(v)| \leq Ci^{-\alpha}$ for some absolute constant $C > 0$, with $\alpha > 1$. Let $\zeta(\cdot)$ denote the Riemann zeta function. Then, for any ℓ, if a maximum margin classifier has a margin $> \Delta$, with probability more than $1 - \delta$, its risk is less than

$$\epsilon := \frac{\tilde{O}\left(\frac{\sqrt{\zeta(2\alpha)\ell}}{\Delta}\right) + \log \frac{1}{\delta}}{\ell},$$

where $\tilde{O}$ hides multiplicative polylogarithmic factors.

This result follows from a bound on the annealed entropy of gap-tolerant classifiers in a Hilbert space that is of independent interest. In addition, it makes important use of the fact that although individual elements of the heavy-tailed feature vector may be large, the vector has bounded moments.

Our second main result provides bounds on classifying data with spectral kernels. In particular, in the following theorem we give dimension-independent upper bounds on the sample complexity of learning a nearly-optimal maximum margin classifier in the feature space of the Diffusion Maps.

Theorem 2 (Spectral Kernels) *Let the following Diffusion map be given:*

$$\Phi : v \mapsto (l_1^k f_1(v), \ldots, l_n^k f_n(v)),$$

where f_i are normalized eigenfunctions (whose $\ell_2(\mu)$) norm is 1, μ being the uniform distribution), l_i are the eigenvalues of the corresponding Markov Chain and $k \geq 0$. Then, for any ℓ, if a maximum margin classifier has a margin $> \Delta$, with probability more than $1 - \delta$, its risk is less than

$$\epsilon := \frac{\tilde{O}\left(\frac{\sqrt{\ell}}{\Delta}\right) + \log \frac{1}{\delta}}{\ell},$$

where $\tilde{O}$ hides multiplicative polylogarithmic factors.

As with the proof of our main heavy-tailed learning result, the proof of our main spectral learning result makes essential use of an upper bound on the annealed entropy of gap-tolerant classifiers. In applying it, we make important use of the fact that although individual elements of the feature vector may fluctuate wildly, the norm of the Diffusion Map feature vector is bounded.

As a side remark, note that we are not viewing the feature map in Theorem 2 as necessarily being either a random variable or requiring knowledge of some marginal distribution—as might be the case if one is generating points in some space according to some distribution; then constructing a graph based on nearest neighbors; and then doing diffusions to construct a feature map. Instead, we are thinking of a data graph in which the data are adversarially presented, *e.g.*, a given social network is presented, and diffusions and/or a feature map is then constructed.

These two theorems provides a dimension-independent (*i.e.*, independent of the size n of the graph and the dimension of the feature space) upper bound on the number of samples needed to learn a maximum margin classifier, under the assumption that a heavy-tailed feature map or the Diffusion Map kernel of some scale is used as the feature map. As mentioned, both proofs (described below in Sections 3.3 and 3.4) proceed by providing a dimension-independent upper bound on the annealed entropy of gap-tolerant classifiers in the relevant feature space, and then appealing to Theorem 5 (in Section 2) relating the annealed entropy to the generalization error. For this bound on the annealed entropy of these gap-tolerant classifiers, we crucially use the fact that $\mathbb{E}_v\|\Phi(v)\|^2$ is bounded, even if $\sup_v \|\Phi(v)\|$ is unbounded as $n \to \infty$. That is, although bounds on the individual entries of the feature map do not appear to be known, we crucially use that there exist nontrivial bounds on the magnitude of the feature vectors. Since this bound is of more general interest, we describe it separately.

1.3 Summary of our main technical contribution

The distribution-dependent ideas that underlie our two main results (in Theorems 1 and 2) can also be used to bound the sample complexity of a classification task more generally under the assumption that the expected value of a norm of the data is bounded, *i.e.*, when the magnitude of the feature vector of the data in some norm has a finite moment. In more detail:

- Let $\mathcal{P}$ be a probability measure on a Hilbert space $\mathcal{H}$, and let $\Delta > 0$. In Theorem 6 (in Section 3.1), we prove that if $\mathbb{E}_{\mathcal{P}}\|x\|^2 = r^2 < \infty$, then the annealed entropy of gap-tolerant classifiers (defined in Section 2) in $\mathcal{H}$ can be upper bounded in terms of a function of r, Δ, and (the number of samples) ℓ, independent of the (possibly infinite) dimension of $\mathcal{H}$.

It should be emphasized that the assumption that the expectation of some moment of the norm of the feature vector is bounded is a *much* weaker condition than the more common assumption that the largest element is bounded, and thus this result is likely of more general interest in dealing with heterogeneous and noisy data. For example, similar ideas have been applied recently to the problem of bounding the sample complexity of learning smooth cuts on a low-dimensional manifold [22].

To establish this result, we use a result (See Lemma 2 in Section 3.2.) that the VC dimension of gap-tolerant classifiers in a Hilbert space when the margin is Δ over a bounded domain such as a ball of radius R is bounded above by $\lfloor R^2/\Delta^2 \rfloor + 1$. Such bounds on the VC dimension of gap-tolerant classifiers have been stated previously by Vapnik [27]. However, in the course of his proof bounding the VC dimension of a gap-tolerant classifier whose margin is Δ over a ball of radius R (See [27], page 353.), Vapnik states, without further justification, that due to symmetry the set of points in a ball that is extremal in the sense of being the hardest to shatter with gap-tolerant classifiers is the regular simplex. Attention has been drawn to this fact by Burges (See [4], footnote 20.), who mentions that a rigorous proof of this fact seems to be absent.

Here, we provide a new proof of the upper bound on the VC dimension of such classifiers without making this assumption. (See Lemma 2 in Section 3.2 and its proof.) Hush and Scovel [12] provide an alternate proof of Vapnik's claim; it is somewhat different than ours, and they do not extend their proof to Banach spaces.

The idea underlying our new proof of this result generalizes to the case when the data need not have compact support and where the margin may be measured with respect to more general norms. In particular, we show that the VC dimension of gap-tolerant classifiers with margin Δ in a ball of radius R in a Banach space of Rademacher type $p \in (1, 2]$ and type constant T is bounded above by $\sim (3TR/\Delta)^{\frac{p}{p-1}}$, and that there exists a Banach space of type p (in fact ℓ_p) for which the VC dimension is bounded below by $(R/\Delta)^{\frac{p}{p-1}}$. (See Lemmas 4 and 5 in Section 4.2.) Using this result, we can also prove bounds for the annealed entropy of gap-tolerant classifiers in a Banach space. (See Theorem 7 in Section 4.3.) In addition to being of interest from a theoretical perspective, this result is of potential interest in cases where modeling the relationship between data elements as a dot product in a Hilbert space is too restrictive, and thus this may be of interest, *e.g.*, when the data are extremely sparse and heavy-tailed.

1.4 Maximum margin classification and ERM with gap-tolerant classifiers

Gap-tolerant classifiers—see Section 2 for more details—are useful, at least theoretically, as a means of implementing structural risk minimization (see, *e.g.*, Appendix A.2 of [4]). With gap-tolerant classifiers, the margin Δ is fixed before hand, and does not depend on the data. See, *e.g.*, [9, 11, 12, 26]. With maximum margin classifiers, on the other hand, the margin is a function of the data. In spite of this difference, the issues that arise in the analysis of these two classifiers are similar. For example, through the fat-shattering dimension, bounds can be obtained for the maximum margin classifier, as shown by Shawe-Taylor *et al.* [26]. Here, we briefly sketch how this is achieved.

Definition 1 *Let $\mathcal{F}$ be a set of real valued functions. We say that a set of points $x_1, \ldots, x_s$ is $\gamma-$shattered by $\mathcal{F}$ if there are real numbers $t_1, \ldots, t_s$ such that for all binary vectors $\mathbf{b} = (b_1, \ldots, b_s)$ and each $i \in [s] = \{1, \ldots, s\}$, there is a function $f_{\mathbf{b}}$ satisfying,*

$$f_{\mathbf{b}}(x_i) = \begin{cases} > t_i + \gamma, & \textit{if } b_i = 1; \\ < t_i - \gamma, & \textit{otherwise}. \end{cases} \tag{1}$$

For each $\gamma > 0$, the fat shattering dimension $\mathrm{fat}_{\mathcal{F}}(\gamma)$ of the set $\mathcal{F}$ is defined to be the size of the largest $\gamma-$shattered set if this is finite; otherwise it is declared to be infinity.

Note that, in this definition, t_i can be different for different i, which is not the case in gap-tolerant classifiers. However, one can incorporate this shift into the feature space by a simple construction. We start with the following definition of a Banach space of type p with type constant T.

Definition 2 (Banach space, type, and type constant) *A Banach space is a complete normed vector space. A Banach space $\mathcal{B}$ is said to have (Rademacher) type p if there exists $T < \infty$ such that for all n and $x_1, \ldots, x_n \in \mathcal{B}$*

$$\mathbb{E}_\epsilon[\| \sum_{i=1}^{n} \epsilon_i x_i \|_{\mathcal{B}}^p] \leq T^p \sum_{i=1}^{n} \|x_i\|_{\mathcal{B}}^p.$$

The smallest T for which the above holds with p equal to the type, is called the type constant of $\mathcal{B}$.

Given a Banach space $\mathcal{B}$ of type p and type constant T, let $\mathcal{B}'$ consist of all tuples (v, c) for $v \in \mathcal{B}$ and $c \in \mathbb{R}$, with the norm

$$\|(v,c)\|_{\mathcal{B}'} := \left(\|v\|^p + |c|^p\right)^{1/p}.$$

Noting that if $\mathcal{B}$ is a Banach space of type p and type constant T (see Sections 4.1 and 4.2), one can easily check that $\mathcal{B}'$ is a Banach space of type p and type constant $\max(T, 1)$.

In our distribution-specific setting, we cannot control the fat-shattering dimension, but we can control the logarithm of the expected value of $2^{\kappa(\text{fat}_{\mathcal{F}}(\gamma))}$ for any constant κ by applying Theorem 7 to $\mathcal{B}'$. As seen from Lemma 3.7 and Corollary 3.8 of the journal version of [26], this is all that is required for obtaining generalization error bounds for maximum margin classification. In the present context, the logarithm of the expected value of the exponential of the fat shattering dimension of linear 1-Lipschitz functions on a random data set of size ℓ taken i.i.d from $\mathcal{P}$ on $\mathcal{B}$ is bounded by the annealed entropy of gap-tolerant classifiers on $\mathcal{B}'$ with respect to the push-forward $\mathcal{P}'$ of the measure $\mathcal{P}$ under the inclusion $\mathcal{B} \hookrightarrow \mathcal{B}'$.

This allows us to state the following theorem, which is an analogue of Theorem 4.17 of the journal version of [26], adapted using Theorem 7 of this paper.

Theorem 3 *Let $\Delta > 0$. Suppose inputs are drawn independently according to a distribution $\mathcal{P}$ be a probability measure on a Banach space $\mathcal{B}$ of type p and type constant T, and $\mathbb{E}_{\mathcal{P}}\|x\|^p = r^p < \infty$. If we succeed in correctly classifying ℓ such inputs by a maximum margin hyperplane of margin Δ, then with confidence $1 - \delta$ the generalization error will be bounded from above by*

$$\epsilon := \frac{\tilde{O}\left(\frac{Tr\ell^{\frac{1}{p}}}{\Delta}\right) + \log\frac{1}{\delta}}{\ell},$$

where $\tilde{O}$ hides multiplicative polylogarithmic factors involving ℓ, T, r and Δ.

Specializing this theorem to a Hilbert space, we have the following theorem as a corollary.

Theorem 4 *Let $\Delta > 0$. Suppose inputs are drawn independently according to a distribution $\mathcal{P}$ be a probability measure on a Hilbert space $\mathcal{H}$, and $\mathbb{E}_{\mathcal{P}}\|x\|^2 = r^2 < \infty$. If we succeed in correctly classifying ℓ such inputs by a maximum margin hyperplane with margin Δ, then with confidence $1 - \delta$ the generalization error will be bounded from above by*

$$\epsilon := \frac{\tilde{O}\left(\frac{r\ell^{\frac{1}{2}}}{\Delta}\right) + \log\frac{1}{\delta}}{\ell},$$

where $\tilde{O}$ hides multiplicative polylogarithmic factors involving ℓ, r and Δ.

Note that Theorem 4 is an analogue of Theorem 4.17 of the journal version of [26], but adapted using Theorem 6 of this paper. In particular, note that this theorem does not assume that the distribution is contained in a ball of some radium R, but instead it assumes only that some moment of the distribution is bounded.

1.5 Outline of the paper

In the next section, Section 2, we review some technical preliminaries that we will use in our subsequent analysis. Then, in Section 3, we state and prove our main result for gap-tolerant learning in a Hilbert space, and we show how this result can be used to prove our two main theorems in maximum margin learning. Then, in Section 4, we state and prove an extension of our gap-tolerant learning result to the case when the gap is measured with respect to more general Banach space norms; and then, in Sections 5 and 6 we provide a brief discussion and conclusion. Finally, for completeness, in Appendix A, we will provide a bound for exact (as opposed to maximum margin) learning in the case in which the probability that an entry is nonzero (as opposed to the value of that entry) decays in a heavy-tailed manner.

2 Background and preliminaries

In this paper, we consider the supervised learning problem of binary classification, *i.e.*, we consider an input space $\mathcal{X}$ (*e.g.*, a Euclidean space or a Hilbert space) and an output space $\mathcal{Y}$, where $\mathcal{Y} = \{-1, +1\}$, and where the data consist of pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ that are random variables distributed according to an unknown distribution. We shall assume that for any X, there is at most one pair (X, Y) that is observed. We observe ℓ i.i.d. pairs $(X_i, Y_i), i = 1, \ldots, \ell$ sampled according to this unknown distribution, and the goal is to construct a classification function $\alpha : \mathcal{X} \to \mathcal{Y}$ which predicts $\mathcal{Y}$ from $\mathcal{X}$ with low probability of error.

Whereas an ordinary linear hyperplane classifier consists of an oriented hyperplane, and points are labeled ±1, depending on which side of the hyperplane they lie, a *gap-tolerant classifier* consists of an oriented hyperplane and a margin of thickness Δ in some norm. Any point outside the margin is labeled ±1, depending on which side of the hyperplane it falls on, and all points within the margin are declared "correct," without receiving a ±1 label. This latter setting has been considered in [27, 4] (as a way of implementing structural risk minimization—apply empirical risk minimization to a succession of problems, and choose where the gap Δ that gives the minimum risk bound).

The *risk* $R(\alpha)$ of a linear hyperplane classifier α is the probability that α misclassifies a random data point (x, y) drawn from $\mathcal{P}$; more formally, $R(\alpha) := \mathbb{E}_{\mathcal{P}}[\alpha(x) \neq y]$. Given a set of ℓ labeled data points $(x_1, y_1), \ldots, (x_\ell, y_\ell)$, the *empirical risk* $R_{emp}(\alpha, \ell)$ of a linear hyperplane classifier α is the frequency of misclassification on the empirical data; more formally, $R_{emp}(\alpha, \ell) := \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{I}[x_i \neq y_i]$, where $\mathcal{I}[\cdot]$ denotes the indicator of the respective event. The risk and empirical risk for gap-tolerant classifiers are defined in the same manner. Note, in particular, that data points labeled as "correct" do not contribute to the risk for a gap-tolerant classifier, *i.e.*, data points that are on the "wrong" side of the hyperplane but that are within the Δ margin are not considered as incorrect and do not contribute to the risk.

In classification, the ultimate goal is to find a classifier that minimizes the true risk, *i.e.*, $\arg\min_{\alpha \in \Lambda} R(\alpha)$, but since the true risk $R(\alpha)$ of a classifier α is unknown, an empirical surrogate is often used. In particular, *Empirical Risk Minimization (ERM)* is the procedure of choosing a classifier α from a set of classifiers Λ by minimizing the empirical risk $\arg\min_{\alpha \in \Lambda} R_{emp}(\alpha, \ell)$. The consistency and rate of convergence of ERM—see [27] for precise definitions—can be related to uniform bounds on the difference between the empirical risk and the true risk over all $\alpha \in \Lambda$. There is a large body of literature on sufficient conditions for this kind of uniform convergence. For instance, the VC dimension is commonly-used toward this end. Note that, when considering gap-tolerant classifiers, there is an additional caveat, as one obtains uniform bounds only over those gap-tolerant classifiers that do not contain any data points in the margin—the appendix A.2 of [4] addresses this issue.

In this paper, our main emphasis is on the annealed entropy:

Definition 3 (Annealed Entropy) *Let $\mathcal{P}$ be a probability measure supported on a vector space $\mathcal{H}$. Given a set Λ of decision rules and a set of points $Z = \{z_1, \ldots, z_\ell\} \subset \mathcal{H}$, let $N^{\Lambda}(z_1, \ldots, z_\ell)$ be the number of ways of labeling $\{z_1, \ldots, z_\ell\}$ into positive and negative samples such that there exists* *a gap-tolerant classifier that predicts incorrectly the label of each z_i. Given the above notation,*

$$H^{\Lambda}_{ann}(k) := \ln \mathbb{E}_{\mathcal{P}^{\times k}} N^{\Lambda}(z_1, \ldots, z_k)$$

is the annealed entropy of the classifier Λ with respect to $\mathcal{P}$.

Note that although we have defined the annealed entropy for general decision rules, below we will consider the case that Λ consists of linear decision rules.

As the following theorem states, the annealed entropy of a classifier can be used to get an upper bound on the generalization error. This follows from Theorem 8 in [4] and a remark on page 198 of [8]. Note that the class Λ^* is itself random, and consequently, $\sup_{\alpha \in \Lambda^*} R(\alpha) - R_{emp}(\alpha, \ell)$ is the supremum over a random class.

Theorem 5 *Let Λ^* be the family of all gap-tolerant classifiers such that no data point lies inside the margin. Then,*

$$\mathbb{P}\left[\sup_{\alpha \in \Lambda^*} R(\alpha) - R_{emp}(\alpha, \ell) > \epsilon\right] < 8 \exp\left(\left(H^{\Lambda}_{ann}(\ell)\right) - \frac{\epsilon^2 \ell}{32}\right)$$

holds true, for any number of samples ℓ and for any error parameter ϵ.

The key property of the function class that leads to uniform bounds is the sublinearity of the logarithm of the expected value of the "growth function," which measures the number of distinct ways in which a data set of a particular size can be split by the function class. A finite VC bound guarantees this in a distribution-free setting. The annealed entropy is a distribution-specific measure, *i.e.*, the same family of classifiers can have different annealed entropies when measured with respect to different distributions. For a more detailed exposition of uniform bounds in the context of gap-tolerant classifiers, we refer the reader to ([4], Appendix A.2).

Note also that normed vector spaces (such as Hilbert spaces and Banach spaces) are relevant to learning theory for the following reason. Data are often accompanied with an underlying metric which carries information about how likely it is that two data points have the same label. This makes concrete the intuition that points with the same class label are clustered together. Many algorithms cannot be implemented over an arbitrary metric space, but require a linear structure. If the original metric space does not have such a structure, as is the case when classifying for example, biological data or decision trees, it is customary to construct a feature space representation, which embeds data into a vector space. We will be interested in the commonly-used Hilbert spaces, in which distances in the feature space are measure with respect to the ℓ_2 distance (as well as more general Banach spaces, in Section 4).

Finally, note that our results where the margin is measured in ℓ_2 can be transferred to a setting with kernels. Given a kernel $k(\cdot, \cdot)$, it is well known that linear classification using a kernel $k(\cdot, \cdot)$ is equivalent to mapping x onto the functional $k(x, \cdot)$ and then finding a separating halfspace in the Reproducing Kernel Hilbert Space (RKHS) which is the Hilbert Space generated by the functionals of the form $k(x, \cdot)$. Since the span of any finite set of points in a Hilbert Space can be isometrically embedded in ℓ_2, our results hold in the setting of kernel-based learning as well, when one first uses the feature map $x \mapsto k(x, \cdot)$ and works in the RKHS.

3 Gap-tolerant classifiers in Hilbert spaces

In this section, we state and prove Theorem 6, our main result regarding an upper bound for the annealed entropy of gap-tolerant classifiers in ℓ_2. This result is of independent interest, and it was used in a crucial way in the proof of Theorems 1 and 2. We start in Section 3.1 with the statement and proof of Theorem 6, and then in Section 3.2 we bound the VC dimension of gap-tolerant classifiers over a ball of radius R. Then, in Section 3.3, we apply these results to prove our main theorem on learning with heavy-tailed data, and finally in Section 3.4, we apply these results to prove our main theorem on learning with spectral kernels.

3.1 Bound on the annealed entropy of gap-tolerant classifiers in Hilbert spaces

The following theorem is our main result regarding an upper bound for the annealed entropy of gap-tolerant classifiers. The result holds for gap-tolerant classification in a Hilbert space, *i.e.*, when the distances in the feature space are measured with respect to the ℓ_2 norm. Analogous results hold when distances are measured more generally, as we will describe in Section 4.

Theorem 6 (Annealed entropy; Upper bound; Hilbert Space) *Let $\mathcal{P}$ be a probability measure on a Hilbert space $\mathcal{H}$, and let $\Delta > 0$. If $\mathbb{E}_{\mathcal{P}}\|x\|^2 = r^2 < \infty$, then then the annealed entropy of gap-tolerant classifiers in $\mathcal{H}$, where the gap is Δ, is*

$$H^{\Lambda}_{ann}(\ell) \leq \left(\ell^{\frac{1}{2}}\left(\frac{r}{\Delta}\right)+1\right)(1+\ln(\ell+1)).$$

Proof: Let ℓ independent, identically distributed (i.i.d) samples $z_1, \ldots, z_\ell$ be chosen from $\mathcal{P}$. We partition them into two classes:

$$X = \{x_1, \ldots, x_{\ell-k}\} := \{z_i \mid \|z_i\| > R\},$$

and

$$Y = \{y_1, \ldots, y_k\} := \{z_i \mid \|z_i\| \leq R\}.$$

Our objective is to bound from above the annealed entropy $H^{\Lambda}_{ann}(\ell) = \ln \mathbb{E}[N^{\Lambda}(z_1, \ldots, z_\ell)]$. By Lemma 1, N^{Λ} is sub-multiplicative. Therefore,

$$N^{\Lambda}(z_1, \ldots, z_\ell) \leq N^{\Lambda}(x_1, \ldots, x_{\ell-k})N^{\Lambda}(y_1, \ldots, y_k).$$

Taking an expectation over ℓ i.i.d samples from $\mathcal{P}$,

$$\mathbb{E}[N^{\Lambda}(z_1, \ldots, z_\ell)] \leq \mathbb{E}[N^{\Lambda}(x_1, \ldots, x_{\ell-k})N^{\Lambda}(y_1, \ldots, y_k)].$$

Now applying Lemma 2, we see that

$$\mathbb{E}[N^{\Lambda}(z_1, \ldots, z_\ell)] \leq \mathbb{E}[N^{\Lambda}(x_1, \ldots, x_{\ell-k})(k+1)^{R^2/\Delta^2+1}].$$

Moving $(k+1)^{R^2/\Delta^2+1}$ outside this expression,

$$\mathbb{E}[N^{\Lambda}(z_1, \ldots, z_\ell)] \leq \mathbb{E}[N^{\Lambda}(x_1, \ldots, x_{\ell-k})](k+1)^{R^2/\Delta^2+1}.$$

Note that $N^{\Lambda}(x_1, \ldots, x_{\ell-k})$ is always bounded above by $2^{\ell-k}$ and that the random variables $\mathbb{I}[E_i[\|x_i\| > R]]$ are i.i.d. Let $\rho = \mathbb{P}[\|x_i\| > R]$, and note that $\ell - k$ is the sum of ℓ independent Bernoulli variables. Moreover, by Markov's inequality,

$$\mathbb{P}[\|z_i\| > R] \ \leq \ \frac{\mathbb{E}[\|z_i\|^2]}{R^2},$$

and therefore $\rho \leq (\frac{r}{R})^2$. In addition,

$$\mathbb{E}[N^{\Lambda}(x_1, \ldots, x_{\ell-k})] \leq \mathbb{E}[2^{\ell-k}].$$

Let $I[\cdot]$ denote an indicator variable. $\mathbb{E}[2^{\ell-k}]$ can be written as

$$\prod_{i=1}^{\ell} \mathbb{E}[2^{I[\|z_i\|>R]}] = (1+\rho)^\ell \leq e^{\rho\ell}.$$

Putting everything together, we see that

$$\mathbb{E}[N^\Lambda(z_1,\ldots,z_\ell)] \leq \exp\left(\ell\left(\frac{r}{R}\right)^2 + \ln(\ell+1)\left(\frac{R^2}{\Delta^2}+1\right)\right). \tag{2}$$

If we substitute $R = (\ell r^2\Delta^2)^{\frac{1}{4}}$, it follows that

$$\begin{aligned} H_{ann}^\Lambda(\ell) &= \log \mathbb{E}\left[N^\Lambda(z_1,\ldots,z_\ell)\right] \\ &\leq \left(\ell^{\frac{1}{2}}\left(\frac{r}{\Delta}\right)+1\right)(1+\ln(\ell+1)). \end{aligned}$$

◇

For ease of reference, we note the following easily established fact about N^Λ. This lemma is used in the proof of Theorem 6 above and Theorem 7 below.

Lemma 1 *Let $\{x_1,\ldots,x_\ell\}\cup\{y_1,\ldots,y_k\}$ be a partition of the data Z into two parts. Then, N^Λ is submultiplicative in the following sense:*

$$N^\Lambda(x_1,\ldots,x_\ell,y_1,\ldots y_k) \leq N^\Lambda(x_1,\ldots,x_\ell)N^\Lambda(y_1,\ldots,y_k).$$

Proof: This holds because any partition of $Z := \{x_1,\ldots,x_\ell,y_1,\ldots,y_k\}$ into two parts by an element $\mathcal{I}\in\Lambda$ induces such a partition for the sets $\{x_1,\ldots,x_\ell\}$ and $\{y_1,\ldots,y_\ell\}$, and for any pair of partitions of $\{x_1,\ldots,x_\ell\}$ and $\{y_1,\ldots,y_k\}$, there is at most one partition of Z that induces them.

◇

3.2 Bound on the VC dimension of gap-tolerant classifiers in Hilbert spaces

As an intermediate step in the proof of Theorem 6, we needed a bound on the VC dimension of a gap-tolerant classifier within a ball of fixed radius. Lemma 2 below provides such a bound and is due to Vapnik [27]. Note, though, that in the course of his proof (See [27], page 353.), Vapnik states, without further justification, that due to symmetry the set of points that is extremal in the sense of being the hardest to shatter with gap-tolerant classifiers is the regular simplex. Attention has also been drawn to this fact by Burges ([4], footnote 20), who mentions that a rigorous proof of this fact seems to be absent. Vapnik's claim has since been proved by Hush and Scovel [12]. Here, we provide a new proof of Lemma 2. It is simpler than previous proofs, and in Section 4 we will see that it generalizes to cases when the margin is measured with norms other than ℓ_2.

Lemma 2 (VC Dimension; Upper bound; Hilbert Space) *In a Hilbert-space, the VC dimension of a gap-tolerant classifier whose margin is Δ over a ball of radius R can by bounded above by $\lfloor\frac{R^2}{\Delta^2}\rfloor+1$.*

Proof: Suppose the VC dimension is n. Then there exists a set of n points $X=\{x_1,\ldots,x_n\}$ in $B(R)$ that can be completely shattered using gap-tolerant classifiers. We will consider two cases, first that n is even, and then that n is odd.

First, assume that n is even, i.e., that $n = 2k$ for some positive integer k. We apply the probabilistic method to obtain a upper bound on n. Note that for every set $S \subseteq [n]$, the set $X_S := \{x_i | i \in S\}$ can be separated from $X - X_S$ using a gap-tolerant classifier. Therefore the distance between the centroids (respective centers of mass) of these two sets is greater or equal to 2Δ. In particular, for each S having $k = n/2$ elements,

$$\|\frac{\sum_{i \in S} x_i}{k} - \frac{\sum_{i \not\in S} x_i}{k}\| \geq 2\Delta.$$

Suppose now that S is chosen uniformly at random from the $\binom{n}{k}$ sets of size k. Then,

$$\begin{aligned} 4\Delta^2 &\leq \mathbb{E}\left[\|\frac{\sum_{i \in S} x_i}{k} - \frac{\sum_{i \not\in S} x_i}{k}\|^2\right] \\ &= k^{-2}\left\{\frac{2k+1}{2k}\sum_{i=1}^{n}\|x_i\|^2 - \frac{\|\sum_1^n x_i\|^2}{2k}\right\} \\ &\leq \frac{4(n+1)}{n^2}R^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \Delta^2 &\leq \frac{n+1}{n^2}R^2 \\ &< \frac{R^2}{n-1} \end{aligned}$$

and so

$$n < \frac{R^2}{\Delta^2} + 1.$$

Next, assume that n is odd. We perform a similar calculation for $n = 2k+1$. As before, we average over all sets S of cardinality k the squared distance between the centroid of X_S and the centroid (center of mass) of $X - X_S$. Proceeding as before,

$$\begin{aligned} 4\Delta^2 &\leq \mathbb{E}\left[\|\frac{\sum_{i \in S} x_i}{k} - \frac{\sum_{i \not\in S} x_i}{k+1}\|^2\right] \\ &= \frac{\sum_{i=1}^n \|x_i\|^2(1 + \frac{1}{2n}) - \frac{1}{2n}\|\sum_{1 \leq i \leq n} x_i\|^2}{k(k+1)} \\ &\leq \frac{\sum_{i=1}^n \|x_i\|^2(1 + \frac{1}{2n})}{k(k+1)} \\ &= \frac{4k+3}{2k(2k+1)(k+1)}\{(2k+1)R^2\} \\ &< \frac{4R^2}{n-1}. \end{aligned}$$

Therefore, $n < \frac{R^2}{\Delta^2} + 1$.

◇

3.3 Learning with heavy-tailed data: proof of Theorem 1

Proof: For a random data sample x,

$$\begin{aligned} \mathbb{E}\|\mathbf{x}\|^2 &\leq \sum_{i=1}^{n}(Ci^{-\alpha})^2 & (3) \\ &\leq C^2\zeta(2\alpha), & (4) \end{aligned}$$

where ζ is the Riemann zeta function. The theorem then follows from Theorem 4.

$\diamond$

3.4 Learning with spectral kernels: proof of Theorem 2

Proof: A Diffusion Map for the graph $G = (V, E)$ is the feature map that associates with a vertex x, the feature vector $\mathbf{x} = (l_1^\alpha f_1(x), \ldots, l_m^\alpha f_m(x))$, when the eigenfunctions corresponding to the top m eigenvalues are chosen. Let μ be the uniform distribution on V and $|V| = n$. We note that if the f_j are normalized eigenfunctions, *i.e.*, $\forall j, \sum_{x\in V} f_j(x)^2 = 1$,

$$\mathbb{E}\|\mathbf{x}\|^2 \quad = \quad \frac{\sum_{i=1}^m l_i^{2\alpha}}{n} \leq \frac{\sum_{i=1}^n l_i^{2\alpha}}{n} \leq 1. \tag{5}$$

The above inequality holds because the eigenvalues have magnitudes that are less or equal to 1:

$$1 = l_1 \geq \cdots \geq l_n \geq -1.$$

The theorem then follows from Theorem 4.

$\diamond$

4 Gap-tolerant classifiers in Banach spaces

In this section, we state and prove Theorem 7, our main result regarding an upper bound for the annealed entropy of gap-tolerant classifiers in a Banach space. We start in Section 4.1 with some technical preliminaries; then in Section 4.2 we bound the VC dimension of gap-tolerant classifiers in Banach spaces over a ball of radius R; and finally in Section 4.3 we prove Theorem 7. We include this result for completeness since it is of theoretical interest; since it follows using similar methods to the analogous results for Hilbert spaces that we presented in Section 3; and since this result is of potential practical interest in cases where modeling the relationship between data elements as a dot product in a Hilbert space is too restrictive, *e.g.*, when the data are extremely sparse and heavy-tailed. For recent work in machine learning on Banach spaces, see [7, 21, 20, 28].

4.1 Technical preliminaries

Recall the definition of a Banach space from Definition 2 above. We next state the following form of the Chernoff bound, which we will use in the proof of Lemma 4 below.

Lemma 3 (Chernoff Bound) *Let $X_1, \ldots, X_n$ be discrete independent random variables such that $\mathbb{E}[X_i] = 0$ for all i and $|X_i| \leq 1$ for all i. Let $X = \sum_{i=1}^n X_i$ for all i and σ^2 be the variance of X. Then*

$$\mathbb{P}[|X| \geq \lambda\sigma] \leq 2e^{-\lambda^2/4}$$

for any $0 \leq \lambda \leq 2\sigma$.

4.2 Bounds on the VC dimension of gap-tolerant classifiers in Banach spaces

The idea underlying our new proof of Lemma 2 (of Section 3.2, and that provides an upper bound on the VC dimension of a gap-tolerant classifier in Hilbert spaces) generalizes to the case when the the gap is measured in more general Banach spaces. We state the following lemma for a Banach space of type p with type constant T. Recall, *e.g.*, that ℓ_p for $p \geq 1$ is a Banach space of type $\min(2, p)$ and type constant 1.

Lemma 4 (VC Dimension; Upper bound; Banach Space) *In a Banach Space of type p and type constant T, the VC dimension of a gap-tolerant classifier whose margin is Δ over a ball of radius R can by bounded above by $\left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}} + 64$*

Proof: Since a general Banach space does not possess an inner product, the proof of Lemma 2 needs to be modified here. To circumvent this difficulty, we use Inequality (7) determining the Rademacher type of $\mathcal{B}$. This, while permitting greater generality, provides weaker bounds than previously obtained in the Euclidean case. Note that if $\mu := \frac{1}{n}\sum_{i=1}^{n} x_i$, then by repeated application of the Triangle Inequality,

$$\begin{aligned} \|x_i - \mu\| &\leq (1 - \frac{1}{n})\|x_i\| + \sum_{j \neq i} \frac{\|x_j\|}{n} \\ &< 2 \sup_i \|x_i\|. \end{aligned}$$

This shows that if we start with $x_1, \dots, x_n$ having norm $\leq R$, $\|x_i - \mu\| \leq 2R$ for all i. The property of being shattered by gap-tolerant classifiers is translation invariant. Then, for $\emptyset \subsetneq S \subsetneq [n]$, it can be verified that

$$\begin{aligned} 2\Delta &\leq \left\| \frac{\sum_{i \in S}(x_i - \mu)}{|S|} - \frac{\sum_{i \notin S}(x_i - \mu)}{n - |S|} \right\| \\ &= \frac{n}{2|S|(n - |S|)} \left\| \sum_{i \in S}(x_i - \mu) - \sum_{i \notin S}(x_i - \mu) \right\|. \end{aligned} \tag{6}$$

The Rademacher Inequality states that

$$\mathbb{E}_\epsilon[\|\sum_{i=1}^{n} \epsilon_i x_i\|^p] \leq T^p \sum_{i=1}^{n} \|x_i\|^p. \tag{7}$$

Using the version of Chernoff's bound in Lemma 3

$$\mathbb{P}[|\sum_{i=1}^{n} \epsilon_i| \leq \lambda\sqrt{n}] \geq 1 - 2e^{-\lambda^2/4}. \tag{8}$$

We shall denote the above event by E_λ. Now, let $x_1, \dots, x_n$ be n points in $\mathcal{B}$ with a norm less or equal to R. Let $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$ as before.

$$\begin{aligned} 2^p T^p n R^p &\geq 2^p T^p \sum_{i=1}^{n} \|x_i\|^p \\ &\geq T^p \sum_{i=1}^{n} \|x_i - \mu\|^p \\ &\geq \mathbb{E}_\epsilon[\|\epsilon_i(x_i - \mu)\|^p] \\ &\geq \mathbb{E}_\epsilon[\|\epsilon_i(x_i - \mu)\|^p | E_\lambda] \ \mathbb{P}[E_\lambda] \\ &\geq \mathbb{E}_\epsilon[(n - \lambda^2)^p (2\Delta)^p (1 - 2e^{-\lambda^2/4})] \end{aligned}$$

The last inequality follows from (6) and (8). We infer from the preceding sequence of inequalities that

$$n^{p-1} \leq 2^p T^p \left(\frac{R}{\Delta}\right)^p \left\{(1 - \frac{\lambda^2}{n})^p (1 - 2e^{-\lambda^2/4})\right\}^{-1}.$$

The above is true for any $\lambda \in (0, 2\sqrt{n})$, by the conditions in the Chernoff bound stated in Lemma 3. If $n \geq 64$, choosing λ equal to 8 gives us $n^{p-1} \leq 3^p T^p \left(\frac{R}{\Delta}\right)^p$. Therefore, it is always true that $n \leq \left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}} + 64$.

◇

Finally, for completeness, we next state a lower bound for VC dimension of gap-tolerant classifiers when the margin is measured in a norm that is associated with a Banach space of type $p \in (1, 2]$. Since we are interested only in a lower bound, we consider the special case of ℓ_p^n. Note that this argument does not immediately generalize to Banach spaces of higher type because for $p > 2$, ℓ_p has type 2.

Lemma 5 (VC Dimension; Lower Bound; Banach Space) *For each $p \in (1, 2]$, there exists a Banach space of type p such that the VC dimension of gap-tolerant classifiers with gap Δ over a ball of radius R is greater or equal to*

$$\left(\frac{R}{\Delta}\right)^{\frac{p}{p-1}}.$$

Further, this bound is achieved when the space is ℓ_p.

Proof: We shall show that the first n unit norm basis vectors in the canonical basis can be shattered using gap-tolerant classifiers, where $\Delta = n^{\frac{1-p}{p}}$. Therefore in this case, the VC dimension is $\geq (\frac{R}{\Delta})^{\frac{p}{p-1}}$. Let e_j be the j^{th} basis vector. In order to prove that the set $\{e_1, \dots, e_n\}$ is shattered, due to symmetry under permutations, it suffices to prove that for each k, $\{e_1, \dots, e_k\}$ can be separated from $\{e_{k+1}, \dots, e_n\}$ using a gap-tolerant classifier. Points in ℓ_p are infinite sequences $(x_1, \dots)$ of finite ℓ_p norm. Consider the hyperplane H defined by $\sum_{i=1}^{k} x_i - \sum_{i=k+1}^{n} x_i = 0$. Clearly, it separates the sets in question. We may assume e_j to be e_1, replacing if necessary, k by $n - k$. Let $x = \inf_{y \in H} \|e_1 - y\|_p$. Clearly, all coordinates $x_{n+1}, \dots$ of x are 0. In order to get a lower bound on the ℓ_p distance, we use the power-mean inequality: If $p \geq 1$, and $x_1, \dots, x_n \in \mathbb{R}$,

$$\left(\frac{\sum_{i=1}^{n} |x_i|^p}{n}\right)^{\frac{1}{p}} \geq \frac{\sum_{i=1}^{n} |x_i|}{n}.$$

This implies that

$$\begin{aligned}
\|e_1 - x\|_p &\geq n^{\frac{1-p}{p}} \|e_1 - x\|_1 \\
&= n^{\frac{1-p}{p}} \left(|1 - x_1| + \sum_{i=2}^{n} |x_i|\right) \\
&\geq n^{\frac{1-p}{p}} \left(1 - \sum_{i=1}^{k} x_i + \sum_{i=k+1}^{n} x_i\right) \\
&= n^{\frac{1-p}{p}}.
\end{aligned}$$

For $p > 2$, the type of ℓ_p is 2 [15]. Since $\frac{p}{p-1}$ is a decreasing function of p in this regime, we do not recover any useful bounds.

◇

4.3 Bound on the annealed entropy of gap-tolerant classifiers in Banach spaces

The following theorem is our main result regarding an upper bound for the annealed entropy of gap-tolerant classifiers in Banach spaces. Note that the ℓ_2 bound provided by this theorem is slightly weaker than that provided by Theorem 6. Note also that it may seem counter-intuitive that in the case of ℓ_2 (*i.e.*, when we set $\gamma = 2$), the dependence of Δ is Δ^{-1}, which is weaker than in the VC bound, where it is Δ^{-2}. The explanation is that the bound on annealed entropy here depends on the number of samples ℓ, while the VC dimension does not. Therefore, the weaker dependence on Δ is compensated for by a term that in fact tends to ∞ as the number of samples $\ell \to \infty$.

Theorem 7 (Annealed entropy; Upper bound; Banach Space) *Let $\mathcal{P}$ be a probability measure on a Banach space $\mathcal{B}$ of type p and type constant T. Let $\gamma, \Delta > 0$, and let $\eta = \frac{p}{p+\gamma(p-1)}$. If $\mathbb{E}_{\mathcal{P}}\|x\|^\gamma = r^\gamma < \infty$, then the annealed entropy of gap-tolerant classifiers in $\mathcal{B}$, where the gap is Δ, is*

$$H_{ann}^{\Lambda}(\ell) \leq \left(\eta^{-\eta}(1-\eta)^{-1+\eta}\left(\frac{\ell}{\ln(\ell+1)}\left(\frac{3Tr}{\Delta}\right)^{\gamma}\right)^{\eta} + 64\right)\ln(\ell+1).$$

Proof: The proof of this theorem parallels that of Theorem 6, except that here we use Lemma 4 instead of Lemma 2. We include the full proof for completeness. Let ℓ independent, identically distributed (i.i.d) samples $z_1, \dots, z_\ell$ be chosen from $\mathcal{P}$. We partition them into two classes:

$$X = \{x_1, \dots, x_{\ell-k}\} := \{z_i \mid \|z_i\| > R\},$$

and

$$Y = \{y_1, \dots, y_k\} := \{z_i \mid \|z_i\| \leq R\}.$$

Our objective is to bound from above the annealed entropy $H_{ann}^{\Lambda}(\ell) = \ln \mathbb{E}[N^{\Lambda}(z_1, \dots, z_\ell)]$. By Lemma 1, N^{Λ} is sub-multiplicative. Therefore,

$$N^{\Lambda}(z_1, \dots, z_\ell) \leq N^{\Lambda}(x_1, \dots, x_{\ell-k})N^{\Lambda}(y_1, \dots, y_k).$$

Taking an expectation over ℓ i.i.d samples from $\mathcal{P}$,

$$\mathbb{E}[N^{\Lambda}(z_1, \dots, z_\ell)] \leq \mathbb{E}[N^{\Lambda}(x_1, \dots, x_{\ell-k})N^{\Lambda}(y_1, \dots, y_k)].$$

Now applying Lemma 4, we see that

$$\mathbb{E}[N^{\Lambda}(z_1, \dots, z_\ell)] \leq \mathbb{E}[N^{\Lambda}(x_1, \dots, x_{\ell-k})(k+1)^{(3TR/\Delta)^{\frac{p}{p-1}}+64}].$$

Moving $(k+1)^{((2+o(1)TR/\Delta)^{\frac{p}{p-1}})}$ outside this expression,

$$\mathbb{E}[N^{\Lambda}(z_1, \dots, z_\ell)] \leq \mathbb{E}[N^{\Lambda}(x_1, \dots, x_{\ell-k})](k+1)^{(3TR/\Delta)^{\frac{p}{p-1}}+64}.$$

Note that $N^{\Lambda}(x_1, \dots, x_{\ell-k})$ is always bounded above by $2^{\ell-k}$ and that the random variables $\mathbb{I}[E_i[\|x_i\| > R]]$ are i.i.d. Let $\rho = \mathbb{P}[\|x_i\| > R]$, and note that $\ell - k$ is the sum of ℓ independent Bernoulli variables. Moreover, by Markov's inequality,

$$\mathbb{P}[\|z_i\| > R] \;\leq\; \frac{\mathbb{E}[\|z_i\|^\gamma]}{R^\gamma},$$

and therefore $\rho \leq (\frac{r}{R})^\gamma$. In addition,

$$\mathbb{E}[N^{\Lambda}(x_1, \dots, x_{\ell-k})] \leq \mathbb{E}[2^{\ell-k}].$$

Let $I[\cdot]$ denote an indicator variable. $\mathbb{E}[2^{\ell-k}]$ can be written as

$$\prod_{i=1}^{\ell} \mathbb{E}[2^{I[\|z_i\|>R]}] = (1+\rho)^{\ell} \leq e^{\rho\ell}.$$

Putting everything together, we see that

$$\mathbb{E}[N^{\Lambda}(z_1,\ldots,z_\ell)] \leq \exp\left(\ell\left(\frac{r}{R}\right)^{\gamma} + \ln(k+1)\left(64+\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}}\right). \tag{9}$$

By setting $\eta := \frac{p}{\gamma(p-1)+p}$, and adjusting R so that

$$\ell\left(\frac{r}{R}\right)^{\gamma}\eta^{-1} = (1-\eta)^{-1}\ln(\ell+1)\left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}}.$$

We see that

$$\begin{aligned} \ell\left(\frac{r}{R}\right)^{\gamma} + \left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}} &= \left(\ell\left(\frac{r}{R}\right)^{\gamma}\eta^{-1}\right)^{\eta}\left((1-\eta)^{-1}\ln(\ell+1)\left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}}\right)^{1-\eta} \\ &= \eta^{-\eta}(1-\eta)^{-1+\eta}\left(\ell\left(\frac{3Tr}{\Delta}\right)^{\gamma}\right)^{\eta}. \end{aligned}$$

Thus, it follows that

$$\begin{aligned} H_{ann}^{\Lambda}(\ell) &= \log\mathbb{E}\left[N^{\Lambda}(z_1,\ldots,z_\ell)\right] \\ &\leq \left(\eta^{-\eta}(1-\eta)^{-1+\eta}\left(\frac{\ell}{\ln(\ell+1)}\left(\frac{3Tr}{\Delta}\right)^{\gamma}\right)^{\eta} + 64\right)\ln(\ell+1). \end{aligned}$$

$\diamond$

5 Discussion

In recent years, there has been a considerable amount of somewhat-related technical work in a variety of settings in machine learning. Thus, in this section we will briefly describe some of the more technical components of our results in light of the existing related literature.

- Techniques based on the use of Rademacher inequalities allow one to obtain bounds without any assumption on the input distribution as long as the feature maps are uniformly bounded. See, *e.g.*, [10, 14, 1, 13]. Viewed from this perspective, our results are interesting because the uniform boundedness assumption is not satisfied in either of the two settings we consider, although those settings are ubiquitous in applications. In the case of heavy-tailed data, the uniform boundedness assumption is not satisfied due to the slow decay of the tail and the large variability of the associated features. In the case of spectral learning, uniform boundedness assumption is not satisfied since for arbitrary graphs one can have localization and thus large variability in the entries of the eigenvectors defining the feature maps. In both case, existing techniques based on Rademacher inequalities or VC dimensions fail to give interesting results, but we show that dimension-independent bounds can be achieved by bounding the annealed entropy.

- A great deal of work has focused on using diffusion-based and spectral-based methods for nonlinear dimensionality reduction and the learning a nonlinear manifold from which the data are assumed to be drawn [25]. These results are very different from the type of learning bounds we consider here. For instance, most of those learning results involve convergence to an hypothesized manifold Laplacian and not of learning process itself, which is what we consider here.

- Work by Bousquet and Elisseeff [3] has focused on establishing generalization bounds based on stability. It is important to note that their results assume a given algorithm and show how the generalization error changes when the data are changed, so they get generalization results for a given algorithm. Our results make no such assumptions about working with a given algorithm.

- Gurvits [10] has used Rademacher complexities to prove upper bounds for the sample complexity of learning bounded linear functionals on ℓ_p balls. The results in that paper can be used to derive an upper bound on the VC dimension of gap-tolerant classifiers with margin Δ in a ball of radius R in a Banach space of Rademacher type $p \in (1, 2]$. Constants were not computed in that paper, therefore our results do not follow. Moreover, our paper contains results on distribution specific bounds which were not considered there. Finally, our paper considers the application of these tools to the practically-important settings of spectral kernels and heavy-tailed data that were not considered there.

6 Conclusion

We have considered two simple machine learning problems motivated by recent work in large-scale data analysis, and we have shown that although traditional distribution-independent methods based on the VC-dimension fail to yield nontrivial sampling complexity bounds, we can use distribution-dependent methods to obtain dimension-independent learning bounds. In both cases, we take advantage of the fact that, although there may be individual data points that are "outlying," in aggregate their effect is not too large. Due to the increased popularity of vector space-based methods (as opposed to more purely combinatorial methods) in machine learning in recent years, coupled with the continued generation of noisy and poorly-structured data, the tools we have introduced are likely promising more generally for understanding the effect of noise and noisy data on popular machine learning tasks.

A Exact learning with heavy-tailed data

In this appendix section, we state and prove a second result for dimension-independent learning from data in which the feature map exhibits a heavy-tailed decay. The heavy-tailed model we consider here is different than that considered in Theorem 1, and thus we are able to prove bounds for exact (as opposed to maximum margin) learning. Nevertheless, the techniques are similar, and thus we include this result in this paper for completeness.

Consider the following toy model for classifying web pages using keywords. One approach to this problem could be to associate with each web page the indicator vector corresponding to all keywords that it contains. The dimension of this feature space is the number of possible keywords, which is typically very large, and empirical evidence indicates that the frequency of words decays in a heavy-tailed manner. Thus the VC dimension of the feature space is very large, and in a distribution-free setting it is not possible to classify data in such a feature space unless the number of samples is of the order of the VC dimension. More generally, one might be interested in a bipartite graph, *e.g.*, an "advertiser-keyword" or "author-to-paper" graph, in which the nodes are the stated entities and the edges represent some sort of "interaction" between the entities, in which case similar issues arise.

Here, we show that if the probability that the i^{th} keyword in the above toy example is present is heavy-tailed as a function of i, then the sample complexity of the binary classification problem is dimension-independent. More precisely, the following theorem provides a dimension-independent (*i.e.*, independent of the size n of the graph and the dimension of the feature space) upper bound on the number of samples needed to learn by ERM, with a given accuracy and confidence, a linear hyperplane that classifies heavy-tailed data into positive and negative labels, under the assumption that the probability of the i^{th} coordinate of a random data point being non-zero is less than $Ci^{-\alpha}$ for some $C > 0, \alpha > 1$. The proof of this result proceeds by providing providing a dimension-independent upper bound on the annealed entropy of the class of linear classifiers in $\mathbb{R}^d$, and then appealing to Theorem 5 relating the annealed entropy to the generalization error.

Remark: Note that although the generalization bound provided by the following theorem seems to be pessimistic in α, the dependence on α is tight, at least as α tends to 1. Clearly, when $\alpha = 1$, the expected number of 1's in a random sample becomes asymptotically equal to $\log n$, where n is the dimension, in which case, we do not expect a sample complexity that is dimension-independent.

Theorem 8 (Bounds for Heavy-Tailed Data) *Let $\mathcal{P}$ be a probability distribution in $\mathbb{R}^d$. Suppose $\mathcal{P}[x_i \neq 0] \leq Ci^{-\alpha}$ for some absolute constant $C > 0$, with $\alpha > 1$. Then, the annealed entropy of ordinary linear hyperplane classifiers is*

$$H^{\Lambda}_{ann}(\ell) \quad \leq \quad \left(\frac{C}{\alpha - 1}\ell^{\frac{1}{\alpha}} + 1\right) \ln \ell \tag{10}$$

Consequently, the minimum number of random samples $\ell = \ell(\epsilon, \delta)$ needed to learn, by ERM, a classifier whose risk differs from the minimum risk $R(\alpha)$ by $< \epsilon\sqrt{R(\alpha)}$ with probability $> 1 - \delta$ is less than or equal to

$$2\left(\frac{4}{\epsilon^2}\left(\frac{C2^{\frac{1}{\alpha}}}{\alpha-1} + \ln\frac{4}{\delta}\right)\right)^{\frac{\alpha}{\alpha-1}} \ln\left(\left(\frac{4}{\epsilon^2}\left(\frac{C2^{\frac{1}{\alpha}}}{\alpha-1} + \ln\frac{4}{\delta}\right)\right)^{\frac{\alpha}{\alpha-1}}\right).$$

Proof: Let the event that a sample $z_i = (z_{i1}, z_{i2}, \dots)$ has a non-zero coordinate $z_{ik'}$ for some $k' > \ell^{1/\alpha}$ be denoted E_i. The probability of this event can be bounded as follows. If $\alpha \neq 1$ and $k = \ell^{1/\alpha}$, then

$$\begin{aligned}
\mathbb{P}[E_i] &= \mathbb{P}[\exists k' > \ell^{1/\alpha}, \text{ such that } z_{ik'} \neq 0] \\
&\leq C \sum_{i=k+1}^{\infty} i^{-\alpha} \\
&\leq \frac{Ck^{-\alpha+1}}{\alpha - 1}.
\end{aligned}$$

We partition the z_i into two classes :

$$X = \{x_1, \dots, x_{\ell-m}\} := \{z_i \text{ such that } E_i \text{ holds }\}$$

and

$$Y = \{y_1, \dots, y_m\} := \{z_i \text{ such that } E_i \text{ does not hold }\}.$$

N^Λ is sub-multiplicative by Lemma 1. Taking an expectation over ℓ i.i.d samples from $\mathcal{P}$,

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \quad \leq \quad \mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-m})N^\Lambda(y_1, \dots, y_m)]$$

The dimension of the span of $\{y_1, \dots, y_m\}$ is at most k, and by a result from VC theory ([27], page 159) we have

$$N^\Lambda(y_1, \dots, y_m) \leq \exp(k \ln(\frac{m}{k}) + 1).$$

Then,

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq \mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-m})em^k].$$

Moving em^k outside this expression,

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq \mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-k})]em^k.$$

Note that $N^\Lambda(x_1, \dots, x_{\ell-k})$ is always bounded above by $2^{\ell-k}$ and that the events $E_1, E_2, \dots$ are independent identically distributed. Let $p = \mathbb{P}[E_i]$, and note that $\ell - k$ is the sum of ℓ independent p-Bernoulli variables. In addition,

$$\mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-k})] \leq \mathbb{E}[2^{\ell-k}],$$

and $\mathbb{E}[2^{\ell-k}]$ can be written as

$$\begin{aligned} \prod_{i=1}^{\ell}(1 + \mathbb{P}[E_i]) \quad &= \quad (1+p)^\ell && (11) \\ &\leq \quad e^{p\ell} && (12) \\ &= \quad e^{\ell\,(\frac{C\,k^{-\alpha+1}}{\alpha-1})}. && (13) \end{aligned}$$

Putting everything together, we see that

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq e(\ell)^k e^{\frac{C\ell\,k^{-a+1}}{\alpha-1}}.$$

Since $k = \ell^{\frac{1}{\alpha}}$, we see that

$$\begin{aligned} H^\Lambda_{ann}(\ell) \quad &= \quad \ln \mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] && (14) \\ &\leq \quad \left(\frac{C}{\alpha-1}\ell^{\frac{1}{\alpha}} + 1\right) \ln(\ell). && (15) \end{aligned}$$

In order to obtain sample complexity bounds, we need to apply Theorem 5 and substitute the above expression for annealed entropy. For the probability that the error of ERM exceeds $\epsilon\sqrt{R(\alpha)}$ to be less than δ (where α is the optimal classifier), it is sufficient that ℓ satisfy

$$4 \exp\left(\frac{C2^{1/\alpha}}{\alpha-1}\ell^{\frac{1-\alpha}{\alpha}} \ln(2\ell) - \epsilon^2/4\right)\ell \leq \delta.$$

For this to be true, it is enough that

$$\frac{\epsilon^2\ell^{1-\frac{1}{\alpha}}}{4} \geq \frac{C2^{\frac{1}{\alpha}}\ln(2\ell)}{\alpha-1} + \ln(4/\delta).$$

A calculation shows that

$$\frac{2\alpha\left(\frac{4}{\epsilon^2}\left(\frac{C2^{\frac{1}{\alpha}}}{\alpha-1} + \ln\frac{4}{\delta}\right)\right)^{\frac{\alpha}{\alpha-1}} \ln\left(\frac{4}{\epsilon^2}\left(\frac{C2^{\frac{1}{\alpha}}}{\alpha-1} + \ln\frac{4}{\delta}\right)\right)}{\alpha-1}$$

is a value of ℓ that satisfies the previous expression.

◇

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2002.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[3] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[4] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[6] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21:5–30, 2006.

[7] R. Der and D. Lee. Large-margin classification in Banach spaces. In *Proc. of International Conference on AI and Statistics (AISTAT)*, pages 91–98, 2007.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[9] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296, 1999.

[10] L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. In *Proceedings of the 8th International Conference on Algorithmic Learning Theory*, pages 352–363, 1997.

[11] M. Hein, O. Bousquet, and B. Schölkopf. Maximal margin classification for metric spaces. *Journal of Computer and System Sciences*, 71(3):333–359, 2005.

[12] D. Hush and C. Scovel. On the VC dimension of bounded margin classifiers. *Machine Learning*, 45(1):33–44, 2001.

[13] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[14] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In E. Gine, D. Mason, and J. Wellner, editors, *High Dimensional Probability II*, pages 443–459. Birkhauser, 2000.

[15] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag, 1991.

[16] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. arXiv:0810.1355, October 2008.

[17] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceedings of the 17th International Conference on World Wide Web*, pages 695–704, 2008.

[18] J. Leskovec, K.J. Lang, and M.W. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 631–640, 2010.

[19] M.W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106:697–702, 2009.

[20] S. Mendelson. Learnability in Hilbert spaces with reproducing kernels. *Journal of Complexity*, 18(1):152–170, 2002.

[21] C. A. Micchelli and M. Pontil. A function representation for learning in Banach spaces. In *Proceedings of the 17th Annual Conference on Learning Theory*, pages 255–269, 2004.

[22] H. Narayanan and P. Niyogi. On the sample complexity of learning smooth cuts on a manifold. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 000–000, 2009.

[23] P. Paschou, E. Ziv, E.G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M.W. Mahoney, and P. Drineas. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, 3:1672–1686, 2007.

[24] S. I. Resnick. *Heavy Tailed Phenomena.* Springer-Verlag, 2007.

[25] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee. Spectral methods for dimensionality reduction. In O. Chapelle, B. Schoelkopf, and A. Zien, editors, *Semisupervised Learning*, pages 293–308. MIT Press, 2006.

[26] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

[27] V.N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, New York, 1998.

[28] H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel Banach spaces for machine learning. *The Journal of Machine Learning Research*, 10:2741–2775, 2009.